



Programación Python para Big Data

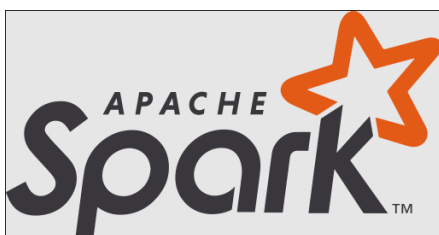
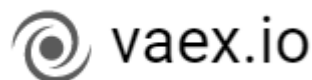
Lección 1: Introducción e Instalaciones

ÍNDICE

Lección 1. – Introducción e Instalaciones	2
Presentación y objetivos	2
1. Contenidos de la Asignatura	5
2. ¿ Qué es considerado Big Data?	7
3. Requisitos previos en Big Data y en IA.....	8
4. Data Engineer Vs Data Scientist	9
5. Instalación de PostgreSQL (no se hará así)	10
6. Instalación de MongoDB (no se hará así)	12
7. Robo 3T para MongoDB (Windows).....	13
8. Robo 3T para MongoDB (Linux).....	16
9. Instalación de Docker y docker-compose (Windows).....	21
10. Instalación de Docker y docker-compose (Linux).....	26
11. Puntos clave.....	33

Lección 1. – Introducción e Instalaciones

PRESENTACIÓN Y OBJETIVOS



Fuentes de obtención de los Logos:

<https://www.python.org/community/logos/>

<https://www.kaggle.com/arunsankar/kaggle-logo>

<https://github.com/scikit-learn/scikit-learn/tree/main/doc/logos>

<https://pandas.pydata.org/about/citing.html>

<https://github.com/numpy/numpy/blob/main/branding/logo/primary/numpylogo.png>

https://commons.wikimedia.org/wiki/File:Jupyter_logo.svg

<https://seaborn.pydata.org/citing.html>

<https://www.mongodb.com/brand-resources>

<https://github.com/pycaret/pycaret/blob/master/logo.png>

<https://www.docker.com/company/newsroom/media-resources>

<https://github.com/kubernetes/kubernetes/blob/master/logo/logo.png>

https://commons.wikimedia.org/wiki/File:Apache_Spark_logo.svg

https://commons.wikimedia.org/wiki/File:Hadoop_logo.svg

<https://vaex.io/>

<https://docs.dask.org/en/latest/logos.html>

<https://www.postgresql.org/about/press/presskit12/es/>

<https://qiskit.org/overview/>

Esta asignatura sobre Programación Python para Big Data dispone de 10 horas de docencia para explicar Big Data y Bases de Datos (BBDD) lo cual, a priori se convierte en una compleja tarea, no obstante, se tratará de hacer un amplio contenido, acorde a la materia.

El planteamiento inicial es que dentro del Big Data se puede enseñar alguna herramienta popular y que se lleve usando años, pero, quizá también alguna más reciente, que quizá gane importancia en los próximos meses o años. (Fecha aproximada de creación de contenidos Junio 2021).

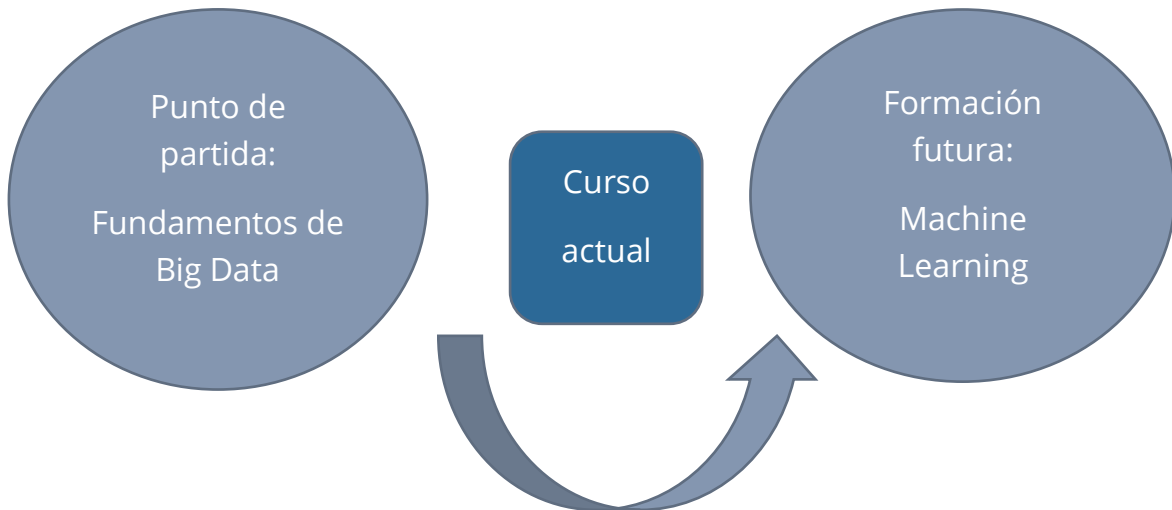


Objetivos

- Introducción a la asignatura y motivos de impartir este temario.
- Guías de instalación de todas las herramientas a utilizar en los 9 temas restantes.

1. CONTENIDOS DE LA ASIGNATURA

Explicación del contenido global del curso y de los motivos de estar así enfocado



Se hace difícil explicar Big Data sin explicar algo de Machine Learning, por lo que no entraremos al detalle del todo, pero explicaremos algunas cosas.

De igual manera, tomaremos como aprendizaje lo visto en "Creación de Aplicaciones Python" y en "Fundamentos de Big Data".

En este caso hablaremos de Pycaret, que es una de las mejores herramientas en Machine Learning (AutoML) y haremos algún ejemplo práctico con varios conceptos importantes.

Trataremos de hacer un contenido balanceado, y adecuado para que el/la estudiante entienda la relación que existe en Data Science entre los diferentes conceptos.

Para ello, y en esta asignatura hemos diseñado el siguiente itinerario formativo:

Algunas herramientas útiles

Tema 2: Docker y Kubernetes

Tema 3: PyCaret y AutoML

Bases de Datos (BBDD)

Tema 4: SQL con PostgreSQL

Tema 5: NoSQL con MongoDB

Herramientas principales

Tema 6: VAEX y DASK

Tema 7: Apache Spark con PySpark [1/2]

Tema 8: Apache Spark con PySpark [2/2] (con Hadoop)

Punto de partida de futuros aprendizajes

Tema 9: Quantum Computing y futuro del Big Data

Tema 10: Certificaciones, Cloud Computing, despliegue de web apps. etc.

2. ¿ QUÉ ES CONSIDERADO BIG DATA?

Antes de comenzar, y dado que no lo hemos mencionado antes trataremos brevemente este tema.

Existen diferentes opiniones, y a nivel particular no lo diferencio así.

En mi caso concreto analizo los sets de datos como “large datasets” (o set de datos muy grandes), o sets de datos “normales”.

Ejemplo, si un set de datos tiene 5000 columnas y 5 millones de filas, tal vez a nivel teórico no sea considerado Big Data pero existe la necesidad de trabajar con esa información de forma algo diferente.

Usaremos diferentes técnicas, por su elevado número de columnas.

A su vez, y dicho sea de paso, tal vez usaremos técnicas de reducción de dimensionalidad como PCA o LDA. Y también, tal vez, trataremos de ver cuáles de esas columnas tienen más peso.

Si son 5 millones de filas, tal vez estamos en un punto que no influye tanto la herramienta de procesamiento, pero existen sets de datos con varios cientos de millones, ahí tal vez deberíamos pensar en usar otro tipo de herramientas, las cuales sean más rápidas para que podamos leerlo rápido.

En esta asignatura quizá trabajemos con un set de datos de unos 600 MB, estamos aun tratando de identificar uno lo suficientemente grande para poder ver la diferencia con lo anterior, si bien es cierto que quizá no es, incluso así considerado Big Data.

Según diferentes opiniones Big Data podría ser considerado a partir de aproximadamente 30 TB, aunque este datos a veces es diferente según quien lo comente. En muchos sitios todo lo superior a 500MB - 1TB son considerados Big Data.

Nota:

- 1 KiloByte son 1000 Bytes
- 1 MegaByte son 1 Millón de Bytes
- 1 GygaByte (GB) son 10 elevado a 9 Bytes (1.000.000.000 Bytes)
- 1 TeraByte (TB) son 10 elevado a 12.

3. REQUISITOS PREVIOS EN BIG DATA Y EN IA

Antes de hablar de Big Data e Inteligencia Artificial (IA) sería muy conveniente hablar de matemáticas, y más concretamente de Cálculo, Álgebra y sobre todo de Estadística y Probabilidad.

En este caso, no ha sido de esta forma, pero si hiciera falta tratará de explicarse algo sobre los propios ejemplos.

Y, probablemente exista la oportunidad de aprender algunas cosas en Inteligencia Artificial.

A fecha Julio-Agosto 2021, existe en la industria la necesidad de encontrar perfiles híbridos, perfiles que, aunque sean expertos en una u otra etapa, entiendan todo el proceso de Data Science.

Estos perfiles de Big Data deberán no ser expertos en redes neuronales o en predicción, pero deben entender qué se hace en cada etapa.

El punto principal de un Data Engineer es la Arquitectura y Bases de Datos (BBDD) pero, puede haber cambios, de hecho algunas veces se pide todo esto también sucede en el perfil como Data Scientist (es decir, que sepa MongoDB, Apache Spark, etc etc).

De modo que la estadística en este perfil, por ejemplo, no es tan importante como en el caso de un "*Machine Learning Engineer*", pero, no se debería evitar tener unos conceptos mínimos.

Animamos a los alumnos/as que no tienen esa mínima base a aprender cosas de forma autodidacta. Cosas como: media, desviación típica, mediana, etc.

4. DATA ENGINEER VS DATA SCIENTIST

Al terminar esta asignatura habrán sido vistas 2 asignaturas relacionadas con Big Data, una de ellas, con fundamentos teóricos, vistos muy brevemente y toda la parte de ploteo de gráficos de un modo más holístico, explicando en qué puntos se usan ciertas herramientas, y que momentos se usan otras.

En este caso, lo que se tratará de explicar son las herramientas que permiten trabajar con grandes volúmenes de datos.

En la práctica esto también será tarea de un *Data Scientist*, pero probablemente responsabilidad del *Data Engineer*.

Como se ha comentado en alguna otra ocasión todo dependerá del tamaño de la empresa y de como separen las tareas pero, saber trabajar con grandes volúmenes de datos es necesario para todos los miembros de un equipo de Data Science, no solo para el Experto en Big Data.

A continuación explicaremos algunas guías de instalación y veremos cómo será llevada a cabo esta tarea, qué programas nos hacen falta, etc.

5. INSTALACIÓN DE POSTGRESQL (NO SE HARÁ ASÍ)

A continuación comentaremos cosas relativas a la Instalación de PostgreSQL, pero procura no instalar nada. Pronto comentaremos algo al respecto.

La forma más típica de instalar PostgreSQL sería irnos a la propia web y buscar la instalación que se adapte a nuestro Sistema Operativo.

<https://www.postgresql.org/download/>

Downloads

PostgreSQL Downloads

PostgreSQL is available for download as ready-to-use packages or installers for various platforms, as well as a source code archive if you want to build it yourself.

Packages and Installers

Select your operating system family:

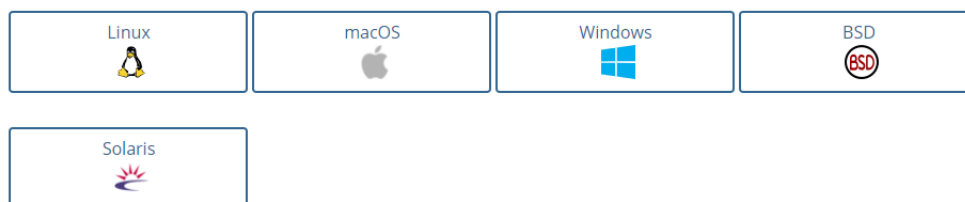


Figura 5.1: Instalación de PostgreSQL

Si eres usuario Linux, puedes elegir el tipo de distribución de igual manera

Select your operating system family:



Select your Linux distribution:

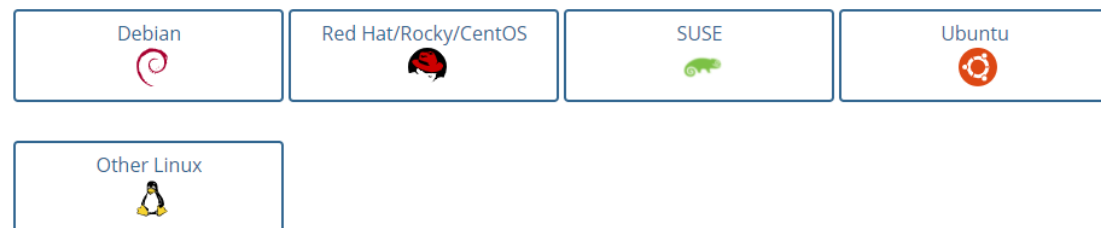


Figura 5.2: Instalación de PostgreSQL

Dado que estamos en una Asignatura de Big Data sería ideal no instalar PostgreSQL de la forma “general”.

Entonces, trata de NO instalar PostgreSQL, pero ten en cuenta que esa es la forma general, no la que seguiremos en la presente Asignatura.

6. INSTALACIÓN DE MONGODB (NO SE HARÁ ASÍ)

A continuación añadiremos una explicación sobre la instalación que deberíamos hacer de MongoDB, pero NO instales nada.

Siendo la presente una Asignatura relacionada con Big Data, no lo haremos de esta forma. Pronto comentaremos cómo haremos esta tarea.

No obstante, la web donde se encuentra tal información es:

<https://www.mongodb.com/es>

En la misma pudiéramos ir a la versión Gratuita.

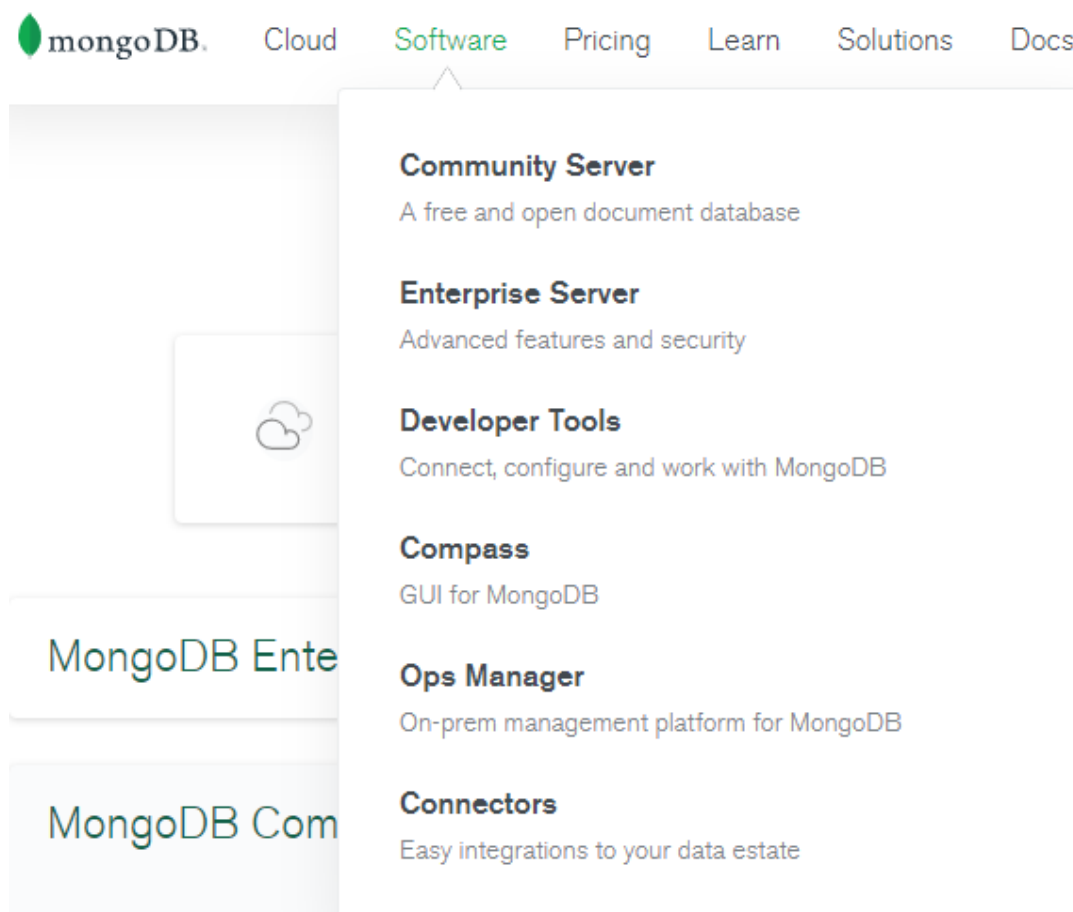


Figura 6.1: Instalación de MongoDB

7. ROBO 3T PARA MONGODB (WINDOWS)

Existe una interfaz que se llama Robomongo, que desde hace años se llama Robo 3T que permite trabajar fácilmente de manera visual con estas bases de datos. Nos iríamos a la opción de Robo 3T que se encuentra a la derecha.

Esta imagen fue realizada a mediados de Julio 2021, por si cuando lo mires aparece de diferente forma.

<https://robomongo.org/download>

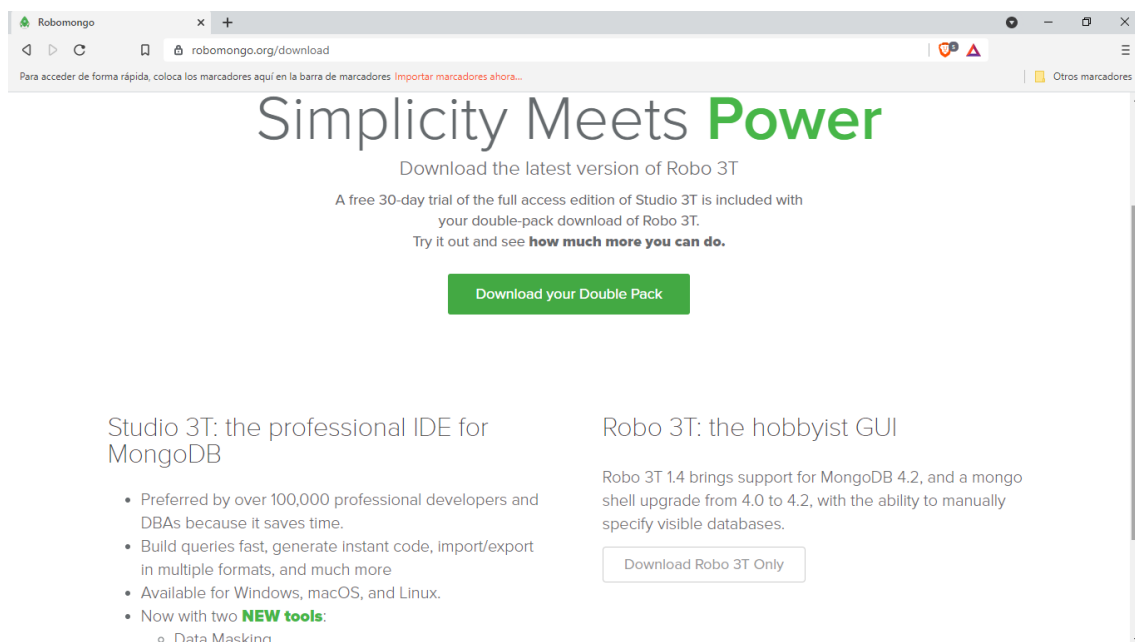


Figura 7.1: Instalación de Robo 3T para MongoDB en Windows (parte 1)

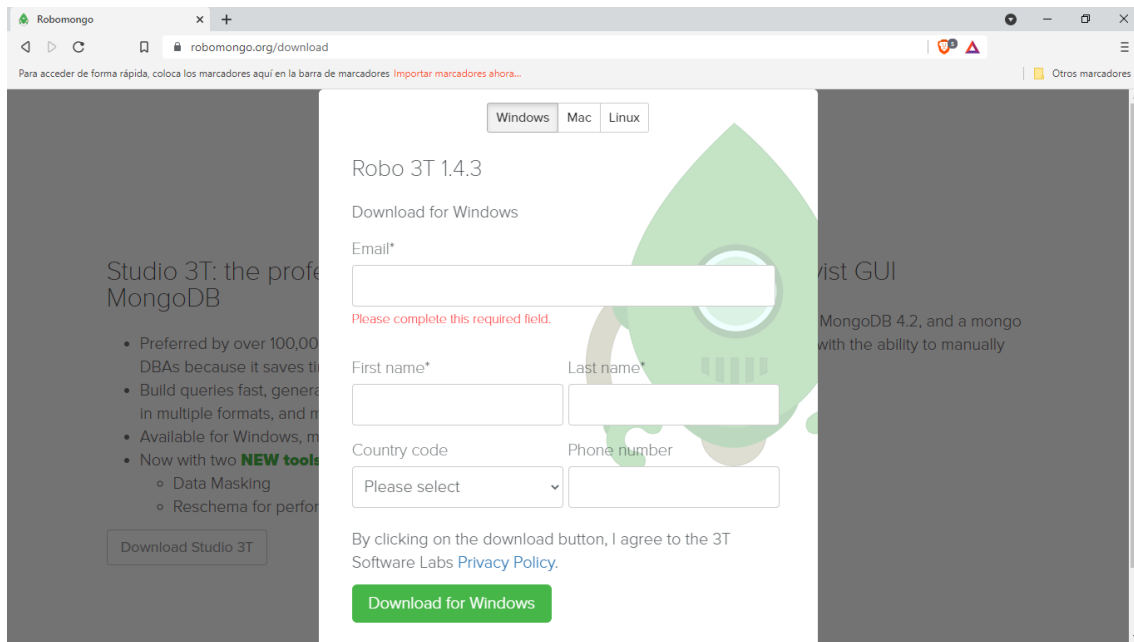


Figura 7.2: Instalación de Robo 3T para MongoDB en Windows (parte 2)

Nos registramos para ello, y en mi caso, Windows, selecciono la opción .exe y le indico donde lo quiero.

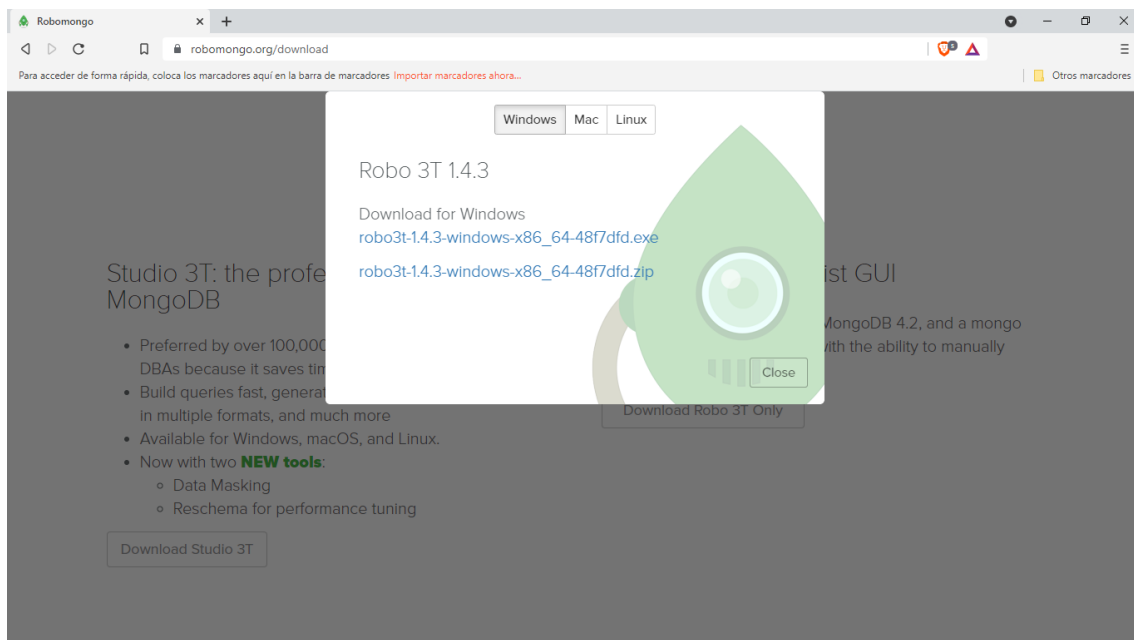


Figura 7.3: Instalación de Robo 3T para MongoDB en Windows (parte 3)

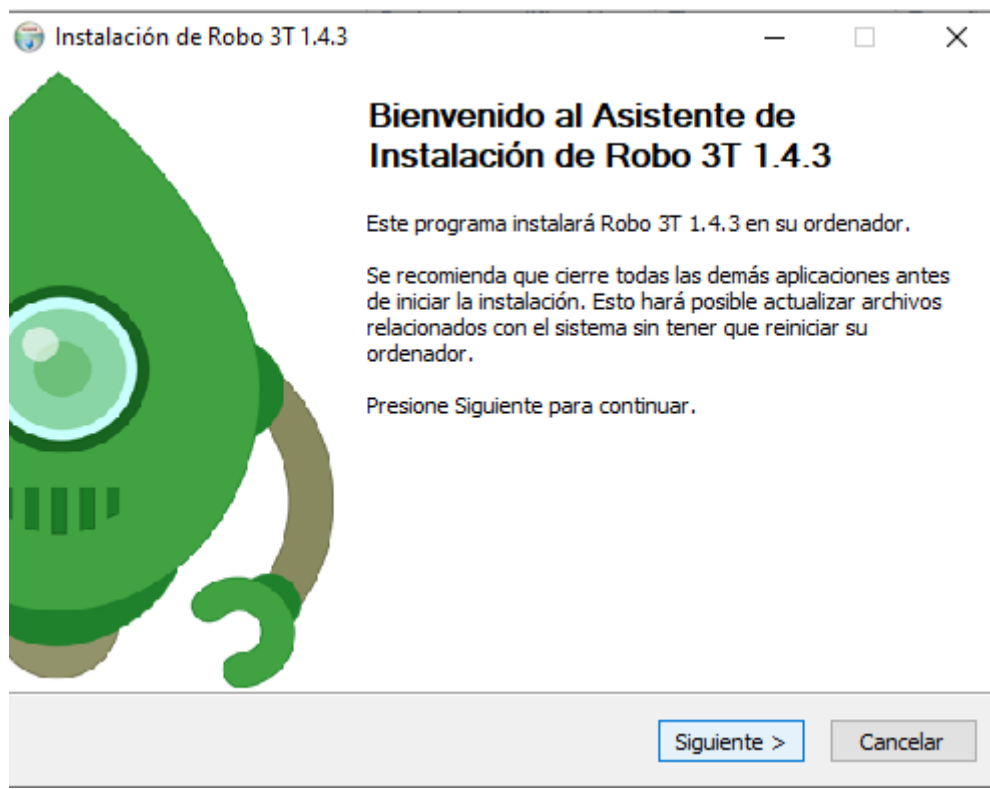


Figura 7.4: Instalación de Robo 3T para MongoDB en Windows (parte 4)

Una vez lo tenga hacemos click sobre el mismo, le indicamos que instale, y será ir indicando siguiente, en un procedimiento muy simple y rápido.

8. ROBO 3T PARA MONGODB (LINUX)

Descargar Robo 3T:

<https://robomongo.org/download>

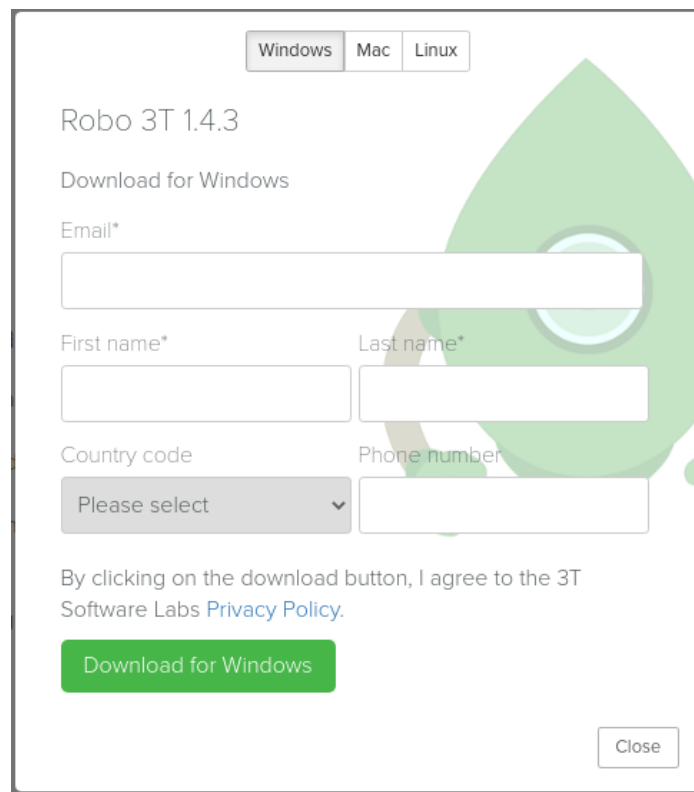
Seleccionar:

Download Robo 3T only

The screenshot shows the Robo 3T download page. At the top, there is a green button labeled "Download your Double Pack". Below this, the page is divided into two columns. The left column is for "Studio 3T: the professional IDE for MongoDB". It lists several bullet points: "Preferred by over 100,000 professional developers and DBAs because it saves time.", "Build queries fast, generate instant code, import/export in multiple formats, and much more", "Available for Windows, macOS, and Linux.", and "Now with two NEW tools: Data Masking and Reschema for performance tuning". Below the list is a button labeled "Download Studio 3T". The right column is for "Robo 3T: the hobbyist GUI". It states "Robo 3T 1.4 brings support for MongoDB 4.2, and a mongo shell upgrade from 4.0 to 4.2, with the ability to manually specify visible databases." Below this text is a button labeled "Download Robo 3T Only", which is highlighted with a red rectangle.

Figura 8.1: Instalación de Robo 3T para MongoDB en Linux (parte 1)

Seleccionar el sistema operativo y cubrir el formulario y descargar:



The screenshot shows a web form for downloading Robo 3T 1.4.3. At the top, there are three tabs: 'Windows', 'Mac', and 'Linux'. The 'Linux' tab is selected. Below the tabs, the text 'Robo 3T 1.4.3' is displayed. Underneath, it says 'Download for Windows'. The form includes several input fields: 'Email*', 'First name*', 'Last name*', 'Country code' (a dropdown menu with 'Please select' as the current selection), and 'Phone number'. A green button labeled 'Download for Windows' is positioned below the form fields. At the bottom right, there is a 'Close' button. A green cartoon robot is visible in the background of the form.

Figura 8.2: Instalación de Robo 3T para MongoDB en Linux (parte 2)

Una vez descargado: descomprimos:

```
tar -xvzf robo3t-1.4.3-linux-x86_64-48f7dfd.tar.gz
```

```
lsabel@lsabel-SVE1512E1EW:~/Documentos$ tar -xvzf robo3t-1.4.3-linux-x86_64-48f7dfd.tar.gz
robo3t-1.4.3-linux-x86_64-48f7dfd/bin/
robo3t-1.4.3-linux-x86_64-48f7dfd/bin/qt.conf
robo3t-1.4.3-linux-x86_64-48f7dfd/bin/robo3t
robo3t-1.4.3-linux-x86_64-48f7dfd/CHANGELOG
robo3t-1.4.3-linux-x86_64-48f7dfd/LICENSE
robo3t-1.4.3-linux-x86_64-48f7dfd/DESCRIPTION
robo3t-1.4.3-linux-x86_64-48f7dfd/COPYRIGHT
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5Gui.so.5.9
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libcrypto.so.1.0.0
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libcui.so.56
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5DBus.la
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5Network.so.5
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5Widgets.prl
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libcui.so
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5XmlPatterns.prl
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libicutest.so.56.1
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5DBus.so.5
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libiculx.so.56.1
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5DBus.prl
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5XmlPatterns.so.5.9.3
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libicudata.so
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libicutu.so
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libcui18n.so
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5XcbQpa.so.5.9.3
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libicule.so.56
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libQt5XcbQpa.prl
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/libicutest.so.56
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/cmake/
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/cmake/GTest/
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/cmake/GTest/GTestTargets.cmake
robo3t-1.4.3-linux-x86_64-48f7dfd/lib/cmake/GTest/GTestConfig.cmake
```

Figura 8.3: Instalación de Robo 3T para MongoDB en Linux (parte 3)

Creamos el archivo robo3t:

```
sudo mkdir /usr/local/bin/robo3t
```

```
lsabel@lsabel-SVE1512E1EW:~/Documentos$ sudo mkdir /usr/local/bin/robo3t
```

Figura 8.4: Instalación de Robo 3T para MongoDB en Linux (parte 4)

Movemos el archivo a la carpeta robo3t:

```
sudo mv robo3t-1.4.3-linux-x86_64-48f7dfd/* /usr/local/bin/robo3t
```

```
isabel@isabel-SVE1512E1EW:~/Documentos$ sudo mv robo3t-1.4.3-linux-x86_64-48f7dfd/* /usr/local/bin/robo3t
```

Figura 8.5: Instalación de Robo 3T para MongoDB en Linux (parte 5)

Como superusuario: `sudo su`

Vamos a la carpeta robo3t:

`cd /usr/local/bin/robo3t/bin`

Si nos muestra permisos denegados:

```
isabel@isabel-SVE1512E1EW:~/Documentos$ cd /usr/local/bin/robo3t/bin
bash: cd: /usr/local/bin/robo3t/bin: Permiso denegado
```

Figura 8.6: Instalación de Robo 3T para MongoDB en Linux (parte 6)

Daremos los permisos a la carpeta:

`sudo chmod 777 /usr/local/bin`

```
isabel@isabel-SVE1512E1EW:~/Documentos$ sudo su
root@isabel-SVE1512E1EW:/home/isabel/Documentos# cd /usr/local/bin/robo3t/bin
root@isabel-SVE1512E1EW:/usr/local/bin/robo3t/bin# sudo chmod +x robo3t ./robo3t
```

Figura 8.7: Instalación de Robo 3T para MongoDB en Linux (parte 7)

Y ejecutaremos para dar permisos de ejecución con la instrucción:

`sudo chmod +x robo3t ./robo3t`

Para ejecutarla :

`cd /usr/local/bin/robo3t/bin`

./robo3t

```
root@isabel-SVE1512E1EW:/usr/local/bin/robo3t/bin# ./robo3t
QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-root'
```

Figura 8.8: Instalación de Robo 3T para MongoDB en Linux (parte 8)

```
isabel@isabel-SVE1512E1EW:/usr/local$ cd /usr/local/bin/robo3t/bin
isabel@isabel-SVE1512E1EW:/usr/local/bin/robo3t/bin$ sudo ./robo3t
QStandardPaths: XDG_RUNTIME_DIR not set, defaulting to '/tmp/runtime-root'
```

Figura 8.9: Instalación de Robo 3T para MongoDB en Linux (parte 9)

Nos muestra la siguiente pantalla:

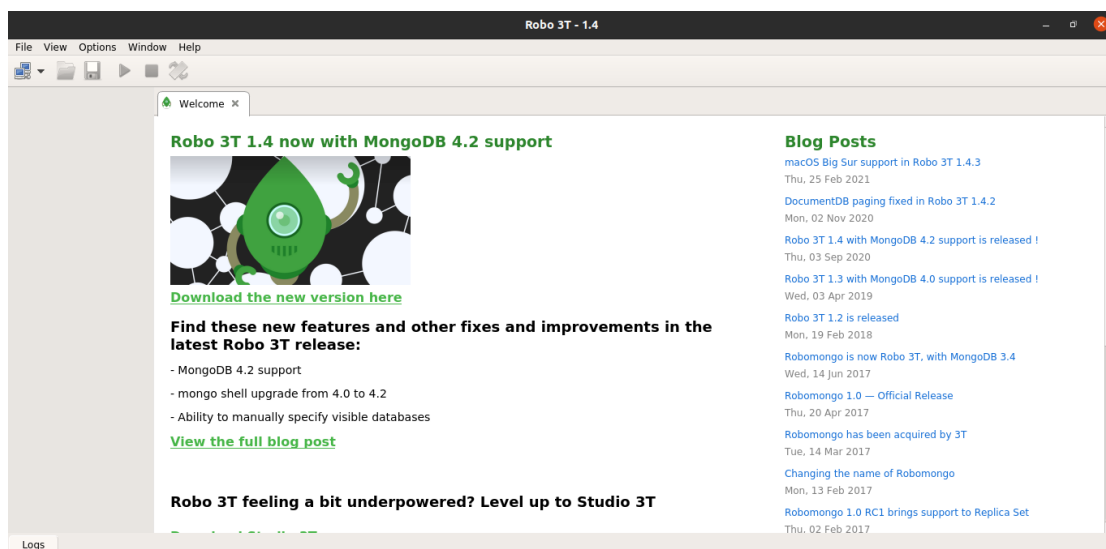


Figura 8.10: Instalación de Robo 3T para MongoDB en Linux (parte 10)

Para conectar con el mongo lo haremos en otro momento.

9. INSTALACIÓN DE DOCKER Y DOCKER-COMPOSE (WINDOWS)

Para instalar Docker en Windows tenemos que irnos al siguiente Link:

<https://docs.docker.com/get-docker/>

y obviamente elegir el sistema operativo correspondiente.

Seleccionamos la opción

Install Docker Desktop on Windows

Estimated reading time: 7 minutes

Welcome to Docker Desktop for Windows. This page contains information about Docker Desktop for Windows system requirements, download URL, installation instructions, and automatic updates.

Docker Desktop for Windows

By downloading Docker Desktop, you agree to the terms of the [Docker Software End User License Agreement](#) and the [Docker Data Processing Agreement](#).

Figura 9.1: Instalación de Docker (parte 1)

Le indicamos la ruta de descarga del ejecutable (de unos 525 MB), y cuando termine le damos doble click para que empiece a instalar.

Nos preguntará algo así como “¿quiere que Docker haga cambios en el equipo? Y le debemos indicar que “sí”.

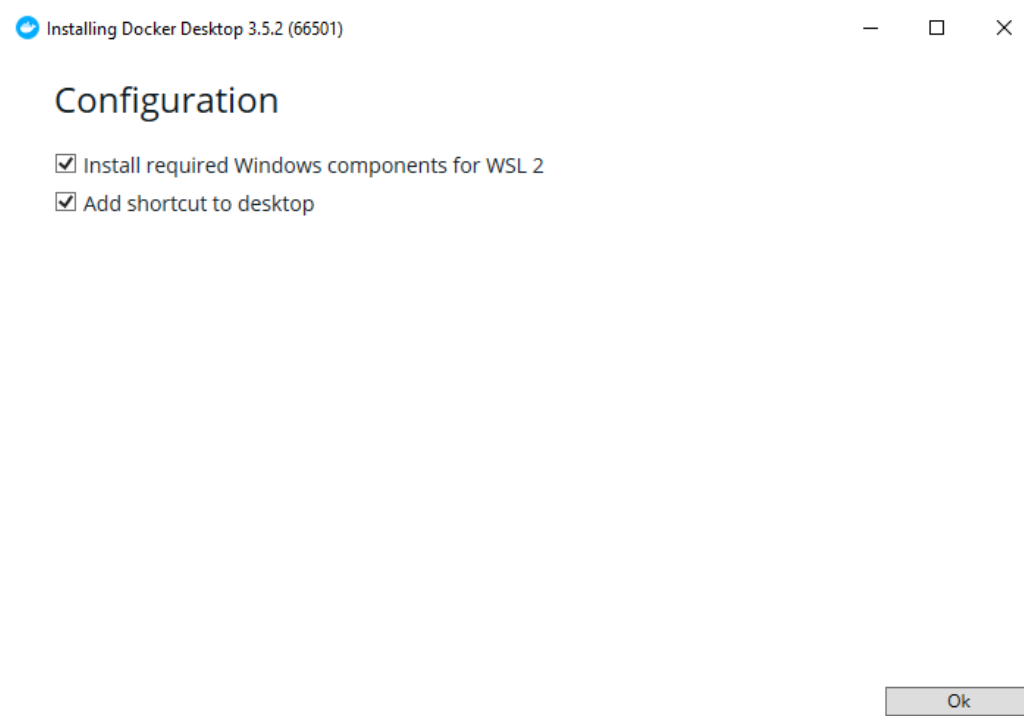


Figura 9.2 Instalación de Docker (parte 2)

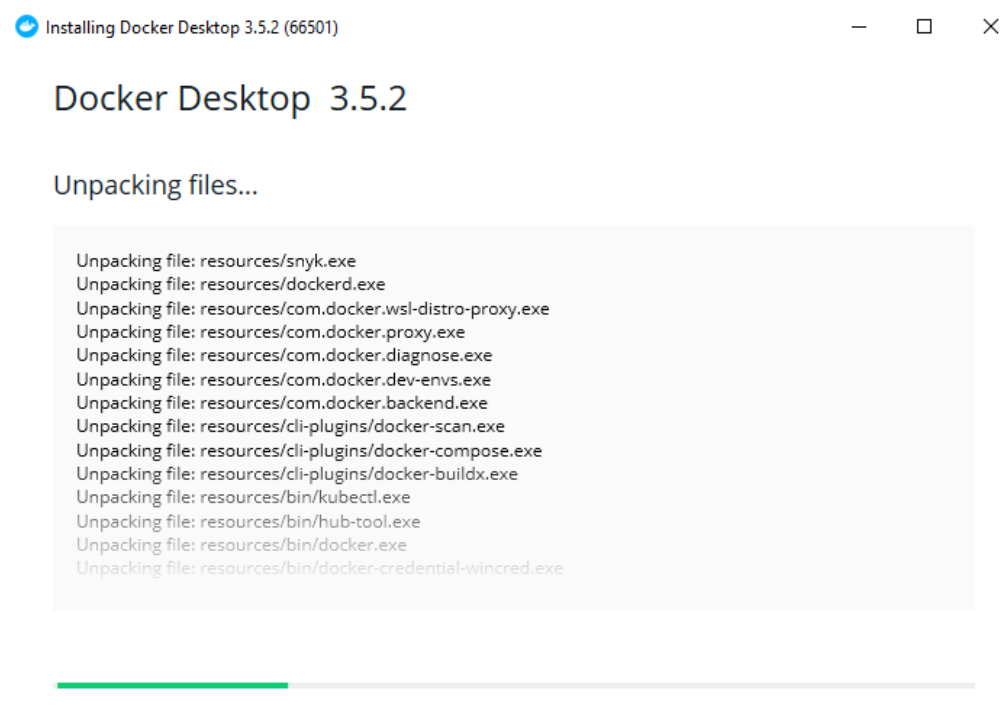


Figura 9.3: Instalación de Docker (parte 3)

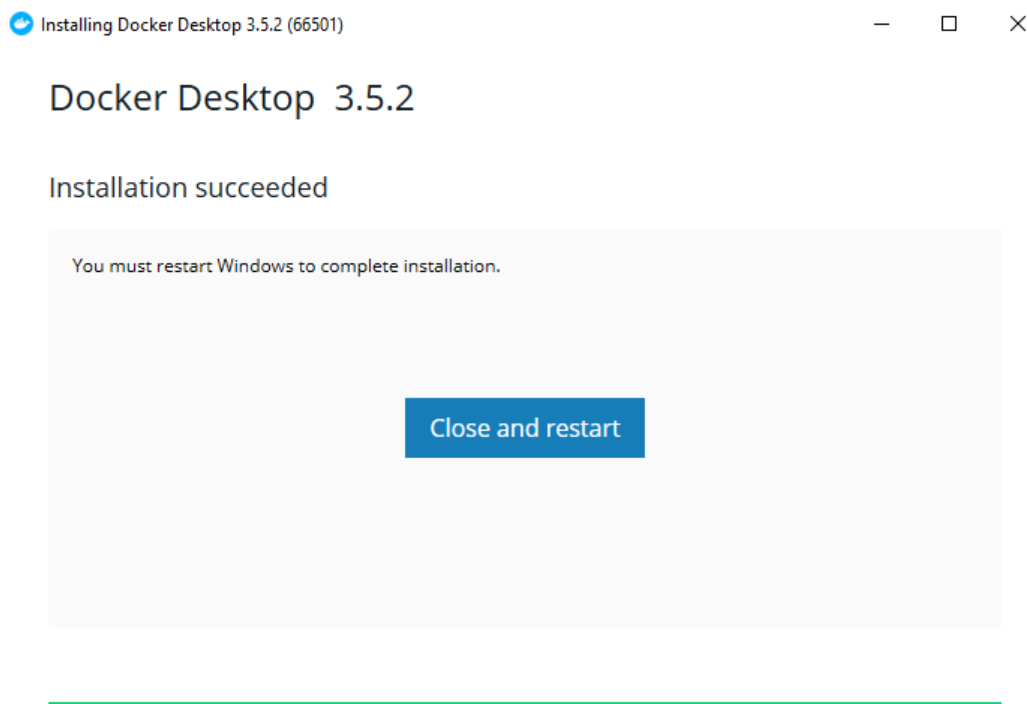


Figura 9.4: Instalación de Docker (parte 4)

Puede que tarde unos minutos hasta que nos ofrezca esta opción..
Reiniciamos el ordenador.

Ahora Docker-compose en Windows ya viene instalado.

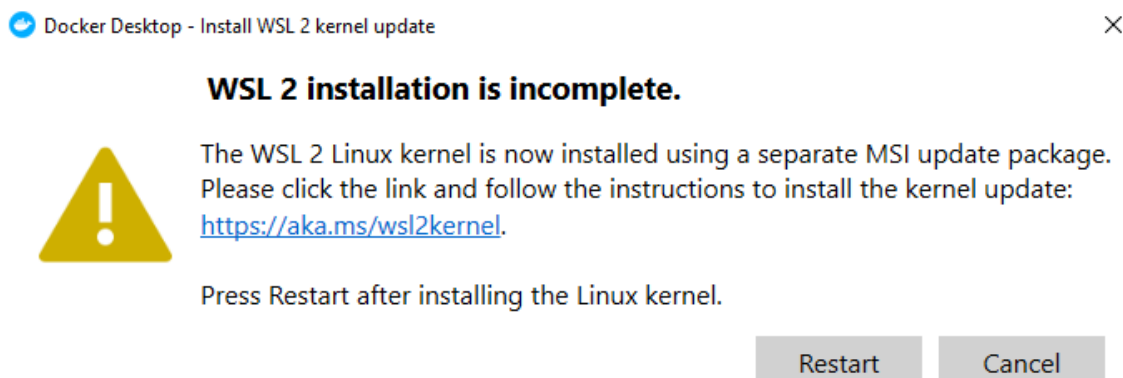


Figura 9.5: Instalación de Docker (parte 5)

Paso 4: Descarga del paquete de actualización del kernel de Linux

1. Descargue la versión más reciente:

- [Paquete de actualización del kernel de Linux en WSL 2 para máquinas x64](#) ↗

Figura 9.6: Instalación de Docker (parte 6)

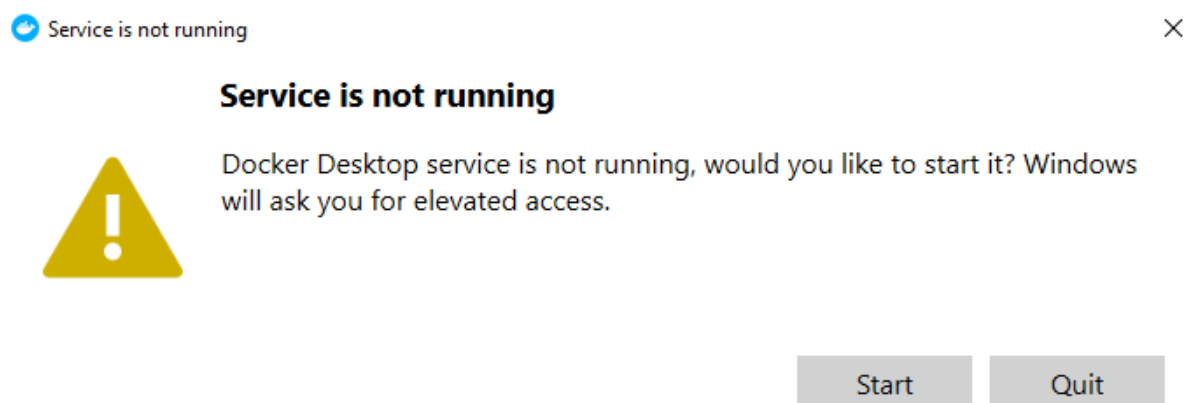


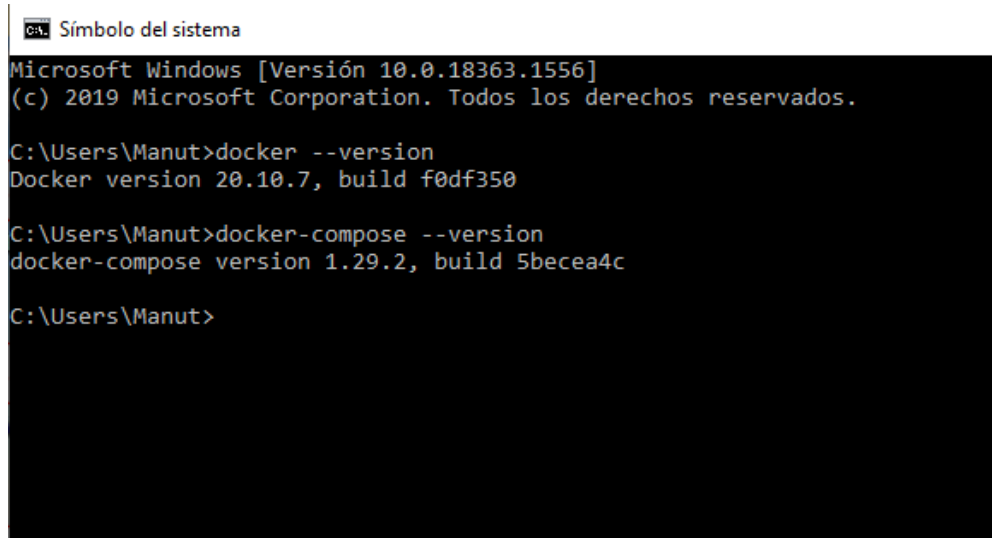
Figura 9.7: Instalación de Docker (parte 7)

Una vez seguidos todos esos pasos tenemos todo instalado.

Ahora solo queda comprobar que es correcto.

Por el momento no vamos a hacer más cosas.

Por ahora, podemos ver la versión que tenemos de la siguiente manera.

A screenshot of a Windows command prompt window. The title bar reads "Símbolo del sistema". The window content shows the following text: "Microsoft Windows [Versión 10.0.18363.1556]", "(c) 2019 Microsoft Corporation. Todos los derechos reservados.", "C:\Users\Manut>docker --version", "Docker version 20.10.7, build f0df350", "C:\Users\Manut>docker-compose --version", "docker-compose version 1.29.2, build 5becea4c", and "C:\Users\Manut>".

```
C:\Users\Manut>docker --version
Docker version 20.10.7, build f0df350

C:\Users\Manut>docker-compose --version
docker-compose version 1.29.2, build 5becea4c

C:\Users\Manut>
```

Figura 9.8: Instalación de Docker (parte 8)

O preferiblemente, incluso: “docker version” y “docker-compose version” y obtendríamos más información.

Aunque por el momento es suficiente.

10. INSTALACIÓN DE DOCKER Y DOCKER-COMPOSE (LINUX)

Para realizar la instalación de docker hay que seguir la guía de instalación facilitada por la página

<https://docs.docker.com/get-docker/>

para los distintos sistemas operativos:

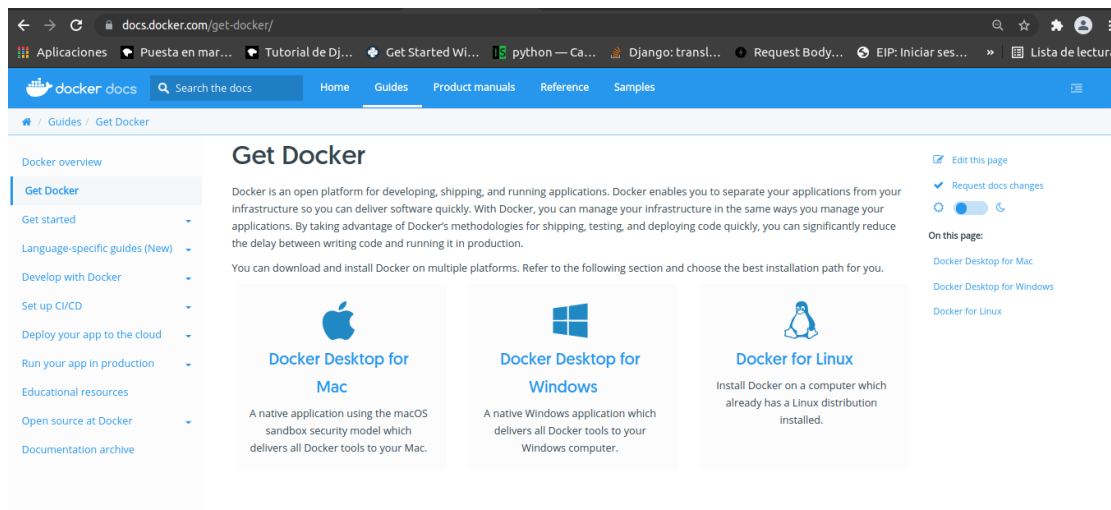


Figura 10.1. Imagen de la guía de instalación en la página oficial Docker

Realizaremos la instalación en mi caso para Ubuntu 20.04:

Instalamos las librerías necesarias:

```
sudo apt-get install apt-transport-https ca-certificates curl gnupg lsb-re
lease
```

```
lsabel@lsabel-SVE1512E1E1:~$ sudo apt-get install apt-transport-https ca-certificates curl gnupg lsb-release
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
lsb-release ya está en su versión más reciente (11.1.0ubuntu2).
fijado lsb-release como instalado manualmente.
ca-certificates ya está en su versión más reciente (20210119~20.04.1).
fijado ca-certificates como instalado manualmente.
gnupg ya está en su versión más reciente (2.2.19-3ubuntu2.1).
fijado gnupg como instalado manualmente.
Se instalarán los siguientes paquetes NUEVOS:
  apt-transport-https curl
0 actualizados, 2 nuevos se instalarán, 0 para eliminar y 24 no actualizados.
Se necesita descargar 163 kB de archivos.
Se utilizarán 571 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] s
Des:1 http://es.archive.ubuntu.com/ubuntu focal-updates/universe amd64 apt-transport-https all 2.0.5 [1.704 B]
Des:2 http://es.archive.ubuntu.com/ubuntu focal-updates/main amd64 curl amd64 7.68.0-1ubuntu2.5 [161 kB]
Descargados 163 kB en 0s (445 kB/s)
Seleccionando el paquete apt-transport-https previamente no seleccionado.
(Leyendo la base de datos ... 201236 ficheros o directorios instalados actualmente.)
Preparando para desempaquetar .../apt-transport-https_2.0.5_all.deb ...
Desempaquetando apt-transport-https (2.0.5) ...
Seleccionando el paquete curl previamente no seleccionado.
Preparando para desempaquetar .../curl_7.68.0-1ubuntu2.5_amd64.deb ...
Desempaquetando curl (7.68.0-1ubuntu2.5) ...
Configurando apt-transport-https (2.0.5) ...
Configurando curl (7.68.0-1ubuntu2.5) ...
Procesando disparadores para man-db (2.9.1-1) ...
```

Figura 10.2. Instalación de librerías

```
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor
or -o /usr/share/keyrings/docker-archive-keyring.gpg
```

```
lsabel@lsabel-SVE1512E1E1:~$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg
```

Figura 10.3. Instalación de librerías

Instrucción para actualización de repositorios:

```
Echo "deb [arch=amd64 signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
| sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
```

```
lsabel@lsabel-SVE1512E1E1:~$ echo "deb [arch=amd64 signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
```

Figura 10.4 Añadir repositorio

Actualización de repositorios:

```
Sudo apt update
```

```
isabel@isabel-SVE1512E1EW:~$ sudo apt-get update
Obj:1 http://es.archive.ubuntu.com/ubuntu focal InRelease
Ign:2 http://dl.google.com/linux/chrome-remote-desktop/deb stable InRelease
Obj:3 http://es.archive.ubuntu.com/ubuntu focal-updates InRelease
Obj:4 http://dl.google.com/linux/chrome/deb stable InRelease
Des:5 https://download.docker.com/linux/ubuntu focal InRelease [52,1 kB]
Obj:6 http://es.archive.ubuntu.com/ubuntu focal-backports InRelease
Obj:7 http://dl.google.com/linux/chrome-remote-desktop/deb stable Release
Des:8 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Des:9 https://download.docker.com/linux/ubuntu focal/stable amd64 Packages [9.960 B]
Des:11 http://security.ubuntu.com/ubuntu focal-security/main amd64 DEP-11 Metadata [24,5 kB]
Des:12 http://security.ubuntu.com/ubuntu focal-security/universe amd64 DEP-11 Metadata [58,4 kB]
Des:13 http://security.ubuntu.com/ubuntu focal-security/multiverse amd64 DEP-11 Metadata [2.468 B]
Descargados 261 kB en 2s (166 kB/s)
```

Figura 10.5 Actualización de repositorios

Instalación de docker:

```
sudo apt-get install docker-ce docker-ce-cli containerd.io
```

```
isabel@isabel-SVE1512E1EW:~$ sudo apt-get install docker-ce docker-ce-cli containerd.io
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
  docker-ce-rootless-extras docker-scan-plugin pigz slirp4netns
Paquetes sugeridos:
  aufs-tools cgroupfs-mount | cgroup-lite
Se instalarán los siguientes paquetes NUEVOS:
  containerd.io docker-ce docker-ce-cli docker-ce-rootless-extras docker-scan-plugin pigz slirp4netns
0 actualizados, 7 nuevos se instalarán, 0 para eliminar y 24 no actualizados.
Se necesita descargar 108 MB de archivos.
Se utilizarán 466 MB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] s
Des:1 https://download.docker.com/linux/ubuntu focal/stable amd64 containerd.io amd64 1.4.6-1 [28,3 MB]
Des:2 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 pigz amd64 2.4-1 [57,4 kB]
Des:3 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 slirp4netns amd64 0.4.3-1 [74,3 kB]
Des:4 https://download.docker.com/linux/ubuntu focal/stable amd64 docker-ce-cli amd64 5:20.10.7~3-0-ubuntu-focal [41,4 MB]
Des:5 https://download.docker.com/linux/ubuntu focal/stable amd64 docker-ce amd64 5:20.10.7~3-0-ubuntu-focal [24,8 MB]
Des:6 https://download.docker.com/linux/ubuntu focal/stable amd64 docker-ce-rootless-extras amd64 5:20.10.7~3-0-ubuntu-focal [9.063 kB]
Des:7 https://download.docker.com/linux/ubuntu focal/stable amd64 docker-scan-plugin amd64 0.8.0-ubuntu-focal [3.889 kB]
Descargados 108 MB en 35s (3.071 kB/s)
Seleccionando el paquete pigz previamente no seleccionado.
(Leyendo la base de datos ... 201247 ficheros o directorios instalados actualmente.)
Preparando para desempaquetar .../0-pigz_2.4-1_amd64.deb ...
Desempaquetando pigz (2.4-1) ...
Seleccionando el paquete containerd.io previamente no seleccionado.
Preparando para desempaquetar .../1-containerd.io_1.4.6-1_amd64.deb ...
Desempaquetando containerd.io (1.4.6-1) ...
Seleccionando el paquete docker-ce-cli previamente no seleccionado.
Preparando para desempaquetar .../2-docker-ce-cli_5%3a20.10.7~3-0-ubuntu-focal_amd64.deb ...
Desempaquetando docker-ce-cli (5:20.10.7~3-0-ubuntu-focal) ...
Seleccionando el paquete docker-ce previamente no seleccionado.
Preparando para desempaquetar .../6-slirp4netns_0.4.3-1_amd64.deb ...
Desempaquetando slirp4netns (0.4.3-1) ...
Configurando slirp4netns (0.4.3-1) ...
Configurando docker-scan-plugin (0.8.0-ubuntu-focal) ...
Configurando containerd.io (1.4.6-1) ...
Created symlink /etc/systemd/system/multi-user.target.wants/containerd.service → /lib/systemd/system/containerd.service.
Configurando docker-ce-cli (5:20.10.7~3-0-ubuntu-focal) ...
Configurando pigz (2.4-1) ...
Configurando docker-ce-rootless-extras (5:20.10.7~3-0-ubuntu-focal) ...
Configurando docker-ce (5:20.10.7~3-0-ubuntu-focal) ...
Created symlink /etc/systemd/system/multi-user.target.wants/docker.service → /lib/systemd/system/docker.service.
Created symlink /etc/systemd/system/sockets.target.wants/docker.socket → /lib/systemd/system/docker.socket.
Procesando disparadores para man-db (2.9.1-1) ...
Procesando disparadores para systemd (245.4-4ubuntu3.6) ...
```

Figura 10.6 Instalación de docker

Comprobación de instalación:

```
sudo docker --version
```

```
isabel@isabel-SVE1512E1EW:~$ sudo docker --version
Docker version 20.10.7, build f0df350
```

Figura 10.7 Versión instalada de docker

Comprobación de si ejecuta correctamente una imagen:

```
sudo docker run hello-world
```

```
isabel@isabel-SVE1512E1EW:~$ sudo docker run hello-world
Unable to find image 'hello-world:latest' locally
latest: Pulling from library/hello-world
b8dfde127a29: Pull complete
Digest: sha256:9f6ad537c5132bcce57f7a0a20e317228d382c3cd61edae14650eec68b2b345c
Status: Downloaded newer image for hello-world:latest

Hello from Docker!
This message shows that your installation appears to be working correctly.

To generate this message, Docker took the following steps:
 1. The Docker client contacted the Docker daemon.
 2. The Docker daemon pulled the "hello-world" image from the Docker Hub.
    (amd64)
 3. The Docker daemon created a new container from that image which runs the
    executable that produces the output you are currently reading.
 4. The Docker daemon streamed that output to the Docker client, which sent it
    to your terminal.

To try something more ambitious, you can run an Ubuntu container with:
$ docker run -it ubuntu bash

Share images, automate workflows, and more with a free Docker ID:
https://hub.docker.com/

For more examples and ideas, visit:
https://docs.docker.com/get-started/
```

Figura 10.8 Ejecución del primer docker de imagen hello-world

Comprobación del servicio activo de docker:

```
sudo systemctl status docker.service
```

```
isabel@isabel-SVE1512E1EW:~$ sudo systemctl status docker.service
● docker.service - Docker Application Container Engine
   Loaded: loaded (/lib/systemd/system/docker.service; enabled; vendor preset: enabled)
   Active: active (running) since Mon 2021-06-21 19:51:07 CEST; 5min ago
     TriggeredBy: ● docker.socket
        Docs: https://docs.docker.com
       Main PID: 13523 (dockerd)
          Tasks: 14
         Memory: 49.6M
        CGroup: /system.slice/docker.service
                └─13523 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/containerd.sock

jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.173976395+02:00" level=warning msg="Your kernel does not support cgroup"
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.173987170+02:00" level=warning msg="Your kernel does not support cgroup"
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.174223561+02:00" level=info msg="Loading containers: start."
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.359661300+02:00" level=info msg="Default bridge (docker0) is assigned with"
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.482407354+02:00" level=info msg="Loading containers: done."
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.549907451+02:00" level=info msg="Docker daemon" commit=b0f5bc3 graphdriver=
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.550227357+02:00" level=info msg="Daemon has completed initialization"
jun 21 19:51:07 isabel-SVE1512E1EW systemd[1]: Started Docker Application Container Engine.
jun 21 19:51:07 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:51:07.588519219+02:00" level=info msg="API listen on /run/docker.sock"
jun 21 19:54:11 isabel-SVE1512E1EW dockerd[13523]: time="2021-06-21T19:54:11.196804628+02:00" level=info msg="ignoring event" container=20fe07f75a97c75
```

Figura 10.9 Comprobación de servicio activo de Docker

Comprobación del servicio activo de containerd:

```
sudo systemctl status containerd.service
```

```
isabel@isabel-SVE1512E1EW:~$ sudo systemctl status containerd.service
● containerd.service - containerd container runtime
   Loaded: loaded (/lib/systemd/system/containerd.service; enabled; vendor preset: enabled)
   Active: active (running) since Mon 2021-06-21 19:51:04 CEST; 6min ago
     Docs: https://containerd.io
       Main PID: 13354 (containerd)
          Tasks: 14
         Memory: 24.7M
        CGroup: /system.slice/containerd.service
                └─13354 /usr/bin/containerd

jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568216979+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568249547+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568269243+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568559549+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568639295+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568721287+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:51:04.568721287+02:00" level=info msg="loading plugin \"/lib/containers/plugins/crio"
jun 21 19:51:04 isabel-SVE1512E1EW systemd[1]: Started containerd container runtime.
jun 21 19:54:10 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:54:10.776300770+02:00" level=info msg="starting signal loop" namespace=moby
jun 21 19:54:11 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:54:11.196783979+02:00" level=info msg="shim disconnected" id=20fe07f75a97c75
jun 21 19:54:11 isabel-SVE1512E1EW containerd[13354]: time="2021-06-21T19:54:11.196849989+02:00" level=error msg="copy shim log" error="read /proc/self
```

Figura 10.10 Comprobación de servicio activo de containerd

Habilitar para que quede siempre activo el docker incluso cuando se reinicie el sistema:

```
sudo systemctl enable containerd.service
```

```
sudo systemctl enable docker.service
```

```
isabel@isabel-SVE1512E1EW:~$ sudo systemctl enable docker.service
Synchronizing state of docker.service with SysV service script with /lib/systemd/systemd-sysv-install.
Executing: /lib/systemd/systemd-sysv-install enable docker
isabel@isabel-SVE1512E1EW:~$ sudo systemctl enable containerd.service
```

Figura 10.11 Habilitar el servicio como activo siempre

Reiniciar, parar o empezar el servicio de docker / containerd:

```
sudo systemctl restart docker.service
```

```
sudo systemctl stop docker.service
```

```
sudo systemctl start docker.service
```

Instalación docker-compose

Podemos ir a la página principal donde encontraremos las distintas instalaciones:

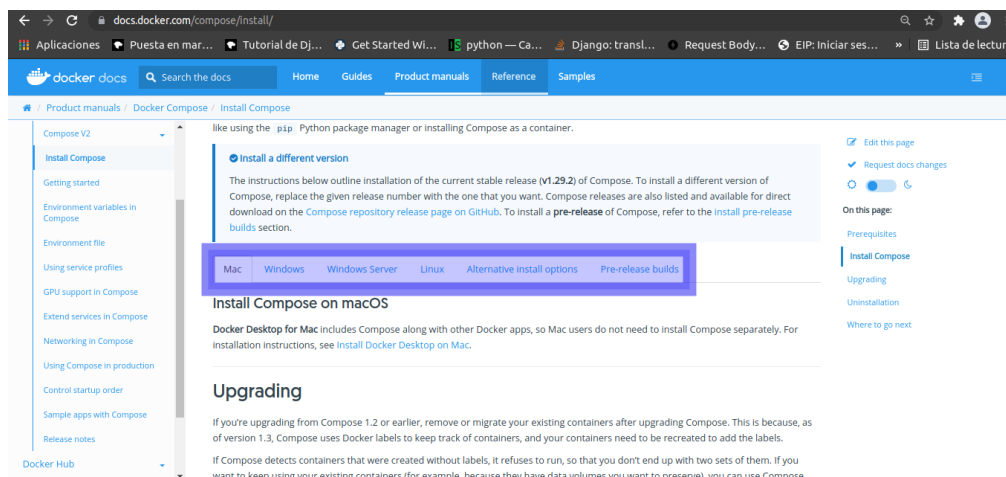


Figura 10.12 Instalación de docker-compose para distintos sistemas operativos

En Ubuntu 20.4 la instalación es simplemente:

```
sudo apt-get install docker-compose
```



```
isabel@isabel-SVE1512E1EW:~/atom/docker$ sudo apt-get install docker-compose
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Los paquetes indicados a continuación se instalaron de forma automática y ya no son necesarios.
  linux-headers-5.8.0-53-generic linux-hwe-5.8.0-53 linux-image-5.8.0-53-generic linux-modules-5.8.0-53-generic
  linux-modules-extra-5.8.0-53-generic
Utilice «sudo apt autoremove» para eliminarlos.
Se instalarán los siguientes paquetes adicionales:
  python3-attr python3-cached-property python3-docker python3-dockerpty python3-dockerpty python3-jsschema python3-pyrsistent python3-texttable
  python3-websocket
Paquetes sugeridos:
  python-attr-doc python-jsschema-doc
Paquetes recomendados:
  docker.io
Se instalarán los siguientes paquetes NUEVOS:
  docker-compose python3-attr python3-cached-property python3-docker python3-dockerpty python3-dockerpty python3-jsschema python3-pyrsistent
0 actualizados, 10 nuevos se instalarán, 0 para eliminar y 29 no actualizados.
Se necesita descargar 391 kB de archivos.
Se utilizarán 2.303 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] S
Des:1 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-cached-property all 1.5.1-4 [10,9 kB]
Des:2 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-websocket all 0.53.0-2ubuntu1 [32,3 kB]
Des:3 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-docker all 4.1.0-1 [83,8 kB]
Des:4 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-dockerpty all 0.4.1-2 [11,1 kB]
Des:5 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-dockerpty all 0.6.2-2.2ubuntu1 [19,7 kB]
Des:6 http://es.archive.ubuntu.com/ubuntu focal/main amd64 python3-attr all 19.3.0-2 [33,9 kB]
Des:7 http://es.archive.ubuntu.com/ubuntu focal/main amd64 python3-pyrsistent amd64 0.15.5-1build1 [52,1 kB]
Des:8 http://es.archive.ubuntu.com/ubuntu focal/main amd64 python3-jsschema all 3.2.0-0ubuntu2 [43,1 kB]
Des:9 http://es.archive.ubuntu.com/ubuntu focal/universe amd64 python3-texttable all 1.6.2-2 [11,0 kB]
Desenpaquetando docker-compose (1.25.0-1) ...
Configurando python3-cached-property (1.5.1-4) ...
Configurando python3-attr (19.3.0-2) ...
Configurando python3-texttable (1.6.2-2) ...
Configurando python3-dockerpty (0.6.2-2.2ubuntu1) ...
Configurando python3-pyrsistent:amd64 (0.15.5-1build1) ...
Configurando python3-websocket (0.53.0-2ubuntu1) ...
update-alternatives: utilizando /usr/bin/python3-wsdump para proveer /usr/bin/wsdump (wsdump) en modo automático
Configurando python3-dockerpty (0.4.1-2) ...
Configurando python3-docker (4.1.0-1) ...
Configurando python3-jsschema (3.2.0-0ubuntu2) ...
Configurando docker-compose (1.25.0-1) ...
Procesando disparadores para man-db (2.9.1-1) ...
```

Figura 10.13 Instalación de docker-compose

Comprobar la instalación, comprobando la versión instalada:

```
sudo docker-compose --version
```

```
isabel@isabel-SVE1512E1EW:~/atom/docker$ sudo docker-compose --version
docker-compose version 1.25.0, build unknown
```

Figura 10.14 Confirmación de instalación de docker-compose

11. PUNTOS CLAVE

- | El temario de una asignatura Big Data pudiera ser mucho más amplio, y a grandes rasgos debe incluir contenido sobre Bases de Datos (BBDD), herramientas para trabajar con grandes sets de datos, conocer herramientas que permiten agilizar trabajo en todo el ciclo de Data Science, y algunas cosas útiles más.
- | A año 2021 las tecnologías que se enseñan en Big Data no pueden ser exactamente las mismas que las que se explicaban con más énfasis en 2014-2018 aproximadamente. Por lo cual se tratará de hacer una combinación de tecnologías consolidadas y herramientas que van a ser importantes en los próximos años.
- | VAEX y Dask son 2 muy buenas elecciones para trabajar con set de datos grandes
- | En Big Data es conveniente trabajar con BBDD relacionales y también No relacionales

